Solution 1: Kullback-Leibler Divergence and model misspecification

(a) The Kullback-Leibler Divergence is defined as:

$$\begin{aligned} D(g, f_\theta) &= \mathbb{E}_g \left[log \left(\frac{g(x)}{f_\theta(x)} \right) \right] \\ &= \underbrace{\mathbb{E}_g \left[log(g(x)) \right]}_{(a)} - \underbrace{\mathbb{E}_g \left[log(f_\theta(x)) \right]}_{(b)} \end{aligned} \tag{1}$$

As we are looking for the set of parameters θ that minimizes $D(g, f_\theta)$, we know the following:

- (a) does not depend on θ, and can be considered as a constant.
- To minimize $D(g, f_\theta)$ is equivalent to maximize (b)

Using the definition of the normal distribution:

$$\begin{aligned} (b) &= \mathbb{E}_g \left[log(f_\theta(x)) \right] \\ &= \mathbb{E}_g \left[log \left(\frac{1}{\sqrt{\sigma^2 2\pi}} \right) - \frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right] \\ &= log \left(\frac{1}{\sqrt{\sigma^2 2\pi}} \right) - \mathbb{E}_g \left[\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right] \\ &= -log\sqrt{\sigma^2 2\pi} - \underbrace{\mathbb{E}_g \left[\frac{1}{2} \frac{x^2 - 2x\mu + \mu^2}{\sigma^2} \right]}_{(c)} \end{aligned} \tag{2}$$

Solving the component (c) in the equation 2 we get:

$$\begin{aligned} (c) &= -\frac{1}{2\sigma^2} \underbrace{\mathbb{E}_g \left[x^2 \right]}_{\mathsf{Var}_g(x) + \mathbb{E}_g[x]^2} + \frac{2\mu}{2\sigma^2} \mathbb{E}_g[x] - \frac{\mu^2}{2\sigma^2} \\ &= -\frac{2\sigma_0^2 + \mu_0^2}{2\sigma^2} + \frac{\mu\mu_0}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \end{aligned} \tag{3}$$

Using the results obtained in 2 and 3, we get the expression that we want to maximize:

$$(b) = -log\sqrt{\sigma^2 2\pi} - \frac{2\sigma_0^2 + \mu_0^2}{2\sigma^2} + \frac{\mu\mu_0}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \tag{4}$$

To maximize 4, we derive the expression with respect to each parameter. We also need to do a second derivative to be sure that the point is a maximum.

First, we derive with respect to the mean parameter μ:

$$\frac{\partial(b)}{\partial\mu} = 0 - 0 + \frac{\mu_0}{\sigma^2} - \frac{\mu}{\sigma^2} \stackrel{!}{=} 0 \longrightarrow \mu_{opt} = \mu_0 \tag{5}$$

This value of μ is a possible maximum, we check the second derivative:

$$\frac{\partial^2(b)}{\partial^2\mu} = -\frac{1}{\sigma^2} < 0 \tag{6}$$

As the second derivative is less than 0 at any point, μ_{opt} maximizes (b) and minimizes the Kullback-Leibler divergence accordingly. We now derive with respect to the variance parameter σ^2:

$$\begin{aligned}\frac{\partial(b)}{\partial\sigma^2} &= -\frac{1}{2\sigma^2} + \frac{2\sigma_0^2+\mu_0^2}{2\sigma^4} - \frac{\mu\mu_0}{\sigma^4} + \frac{\mu^2}{2\sigma^4} \\ &= -\frac{1}{2\sigma^2} + \frac{2\sigma_0^2+\mu_0^2-2\mu\mu_0+\mu^2}{2\sigma^4} \\ &= -\frac{1}{2\sigma^2} + \frac{2\sigma_0^2+(\mu-\mu_0)^2}{2\sigma^4} \overset{!}{=} 0 \longrightarrow \sigma_{opt}^2 = 2\sigma_0^2 + \underbrace{(\mu-\mu_0)^2}_{=0\ if\ \mu=\mu_{opt}}\end{aligned} \tag{7}$$

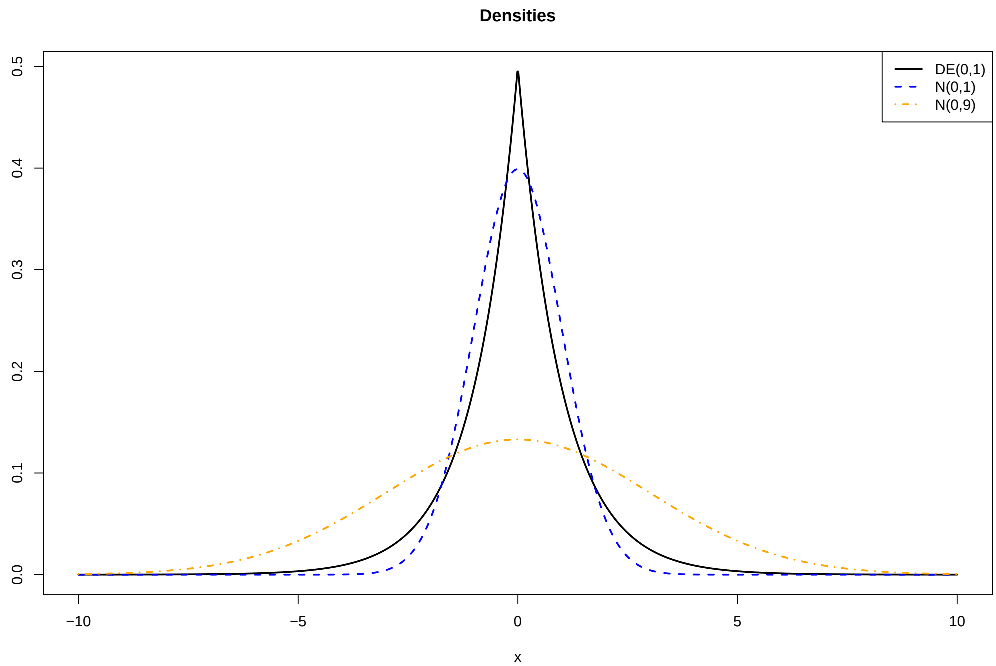
This value of σ^2 is a possible maximum, we check the second derivative:

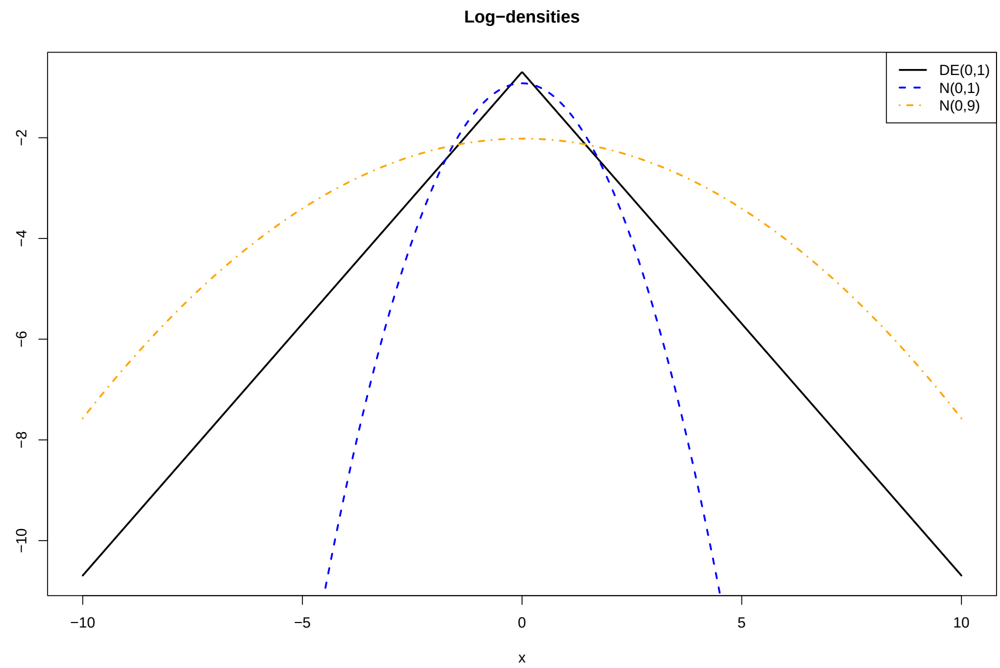
$$\begin{aligned}\frac{\partial^2(b)}{\partial^2\sigma^2} &= \frac{1}{2\sigma^4} - \frac{(2\sigma_0^2+(\mu-\mu_0)^2)}{\sigma^6} \\ \frac{\partial^2(b)}{\partial^2\sigma^2}\Big|_{\sigma^2=\sigma_{opt}^2} &= \frac{1}{2(2\sigma_0^2+(\mu-\mu_0)^2))^2} - \frac{1}{(2\sigma_0^2+(\mu-\mu_0)^2))^2} < 0\end{aligned} \tag{8}$$

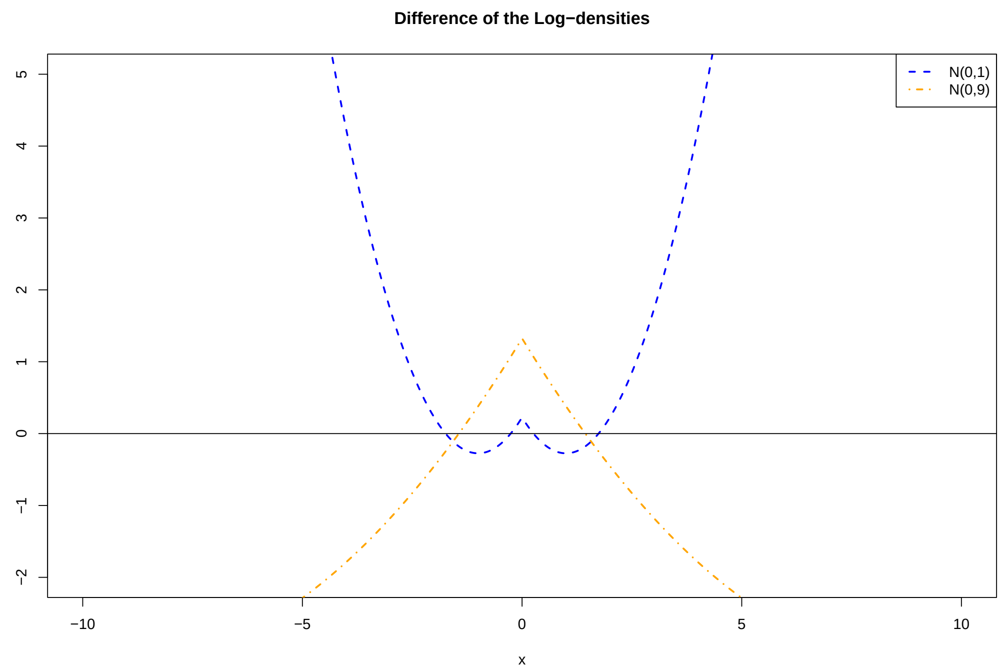
As the second derivative is less than 0 at the point we are looking, σ_{opt}^2 maximizes (b) and thus minimizes the Kullback-Leibler Divergence.

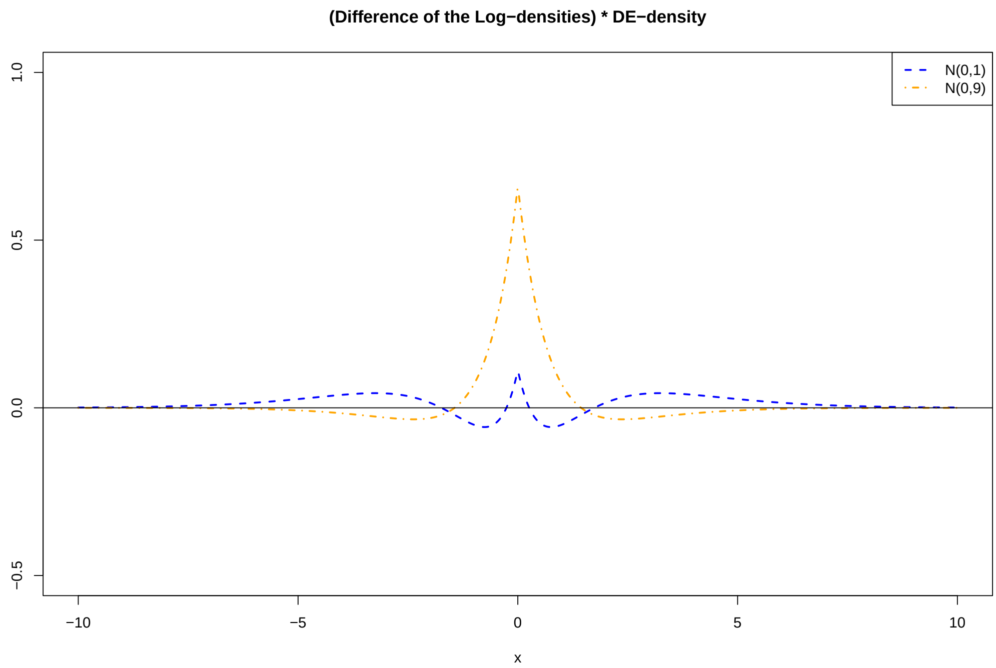
The following graphs may be helpful to understand the problem:

- The KL-Divergence, step by step:



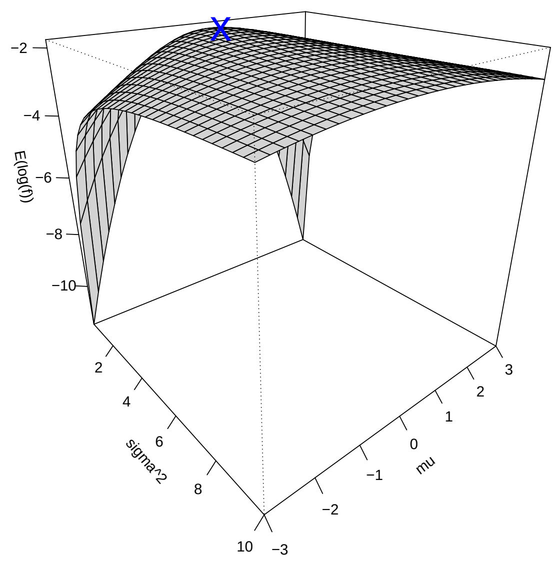

Log−densities
DE(0,1)
N(0,1)
N(0,9)
−2
−4
−6
−8
−10
−10
−5
0
5
10
x

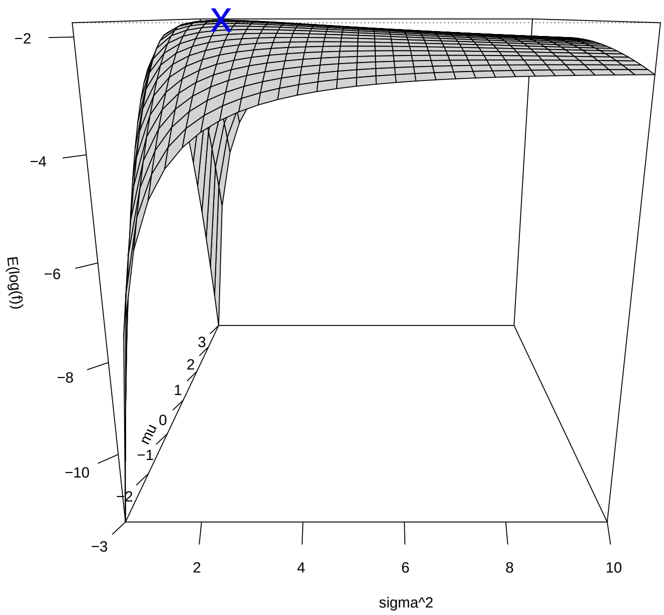

Difference of the Log−densities
N(0,1)
N(0,9)
5
4
3
2
1
0
−1
−2
−10
−5
0
5
10
x

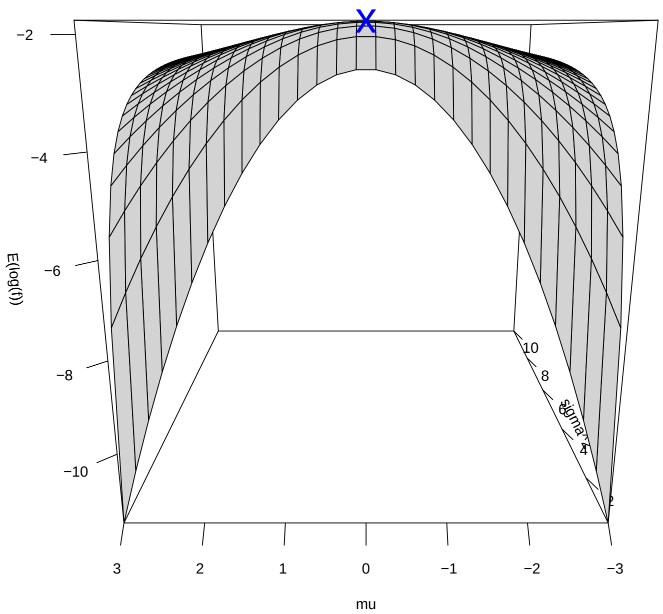


$\Rightarrow$ Now, the KLD is the integral of the function(s) in the last plot.

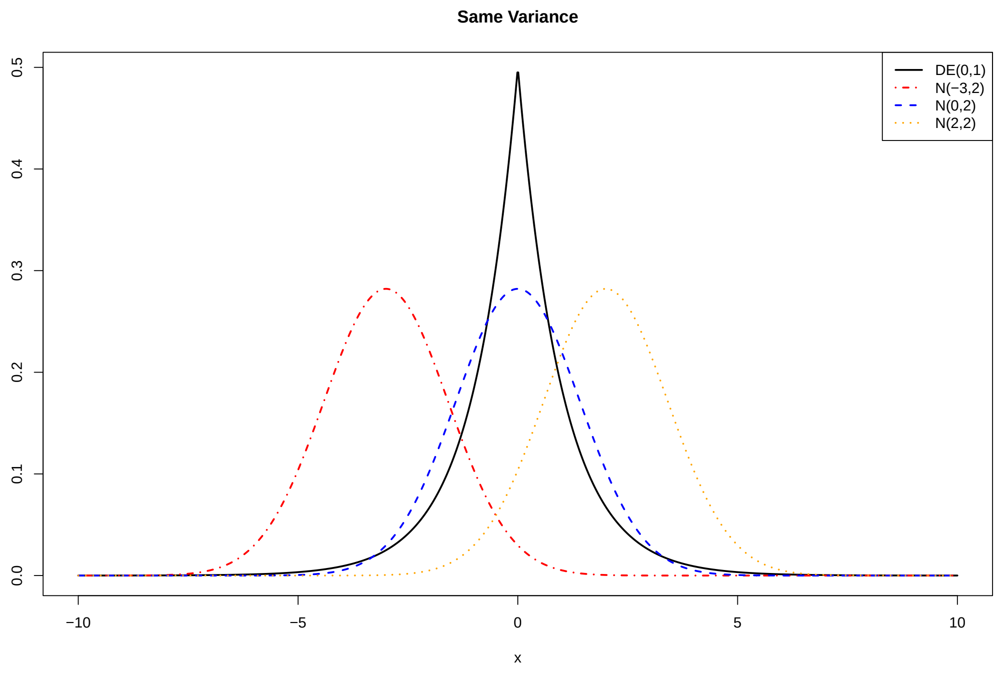
- 3D-Plot of the expression that we want to maximize (for parameters $\mu_0 = 0$ and $\sigma_0 = 1$):

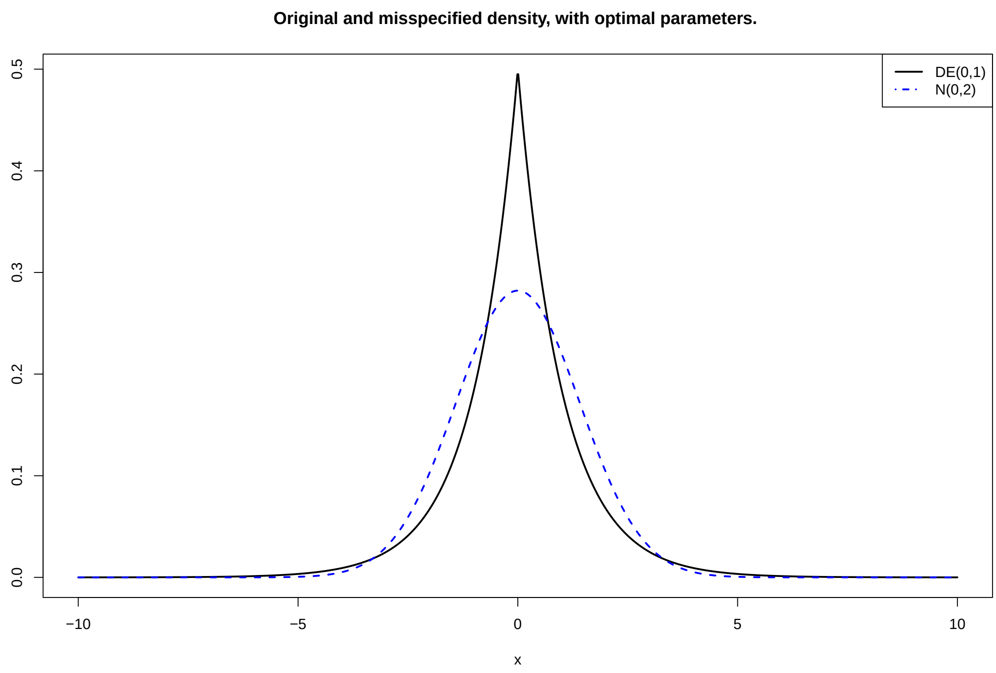






- Different configurations for the parameter μ and the optimal configuration:


Same Variance
DE(0,1)
N(−3,2)
N(0,2)
N(2,2)
0.0
0.1
0.2
0.3
0.4
0.5
−10
−5
0
5
10
x


Original and misspecified density, with optimal parameters.
DE(0,1)
N(0,2)
0.0
0.1
0.2
0.3
0.4
0.5
−10
−5
0
5
10
x